PB86-194412

Studies of Relationships between Molecular Structure and Biological Activity by Pattern Recognition Methods

Pennsylvania State Univ., University Park

Prepared for

Health Effects Research Lab.
Research Triangle Park, NC

1983

U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA/600/D-86/096	2.
	PB86-194412
4. TITLE AND SUBTITLE STUDIES OF RELATIONSHIPS BETWEEN MOLECULAR STRUCTURE AND BIOLOGICAL ACTIVITY BY PATTERN RECOGNITION METHODS	5. REPORT DATE 1983 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) P.C. Jurs	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Pennsylvania State University, University Park, PA 16802	10. PROGRAM ELEMENT NO. 11. CONTRACT/GRANT NO. CR811037
12. SPONSORING AGENCY NAME AND ADDRESS Health Effects Research Laboratory - RTP, NC Office of Research and Development US Environmental Protection Agency Research Triangle Park, NC 27711	13. TYPE OF REPORT AND PERIOD COVERED 14. SPONSORING AGENCY CODE EPA/600/11

15. SUPPLEMENTARY NOTES
Book Chapter - Structure-Activity Correlation as a Predictive Tool in Toxicology, Hemisphere Publishing Co., 1983, Chapter 7, pp. 93-110

16. ABSTRACT

The attempt to rationalize the connections between the molecular structures of organic compounds and their biological activities comprises the field of structure-activity relations (SAR) studies. Correlations between structure and activity are important for the understanding and the development of pharmaceutical agents, agricultural agents, and chemical communicants (olfactory and gustatory stimulants) and for the investigation of chemical and genetic toxicity (mutagenic and carcinogenic potential). Practical importance attaches to this type of study because the potential exists for prediction of the biological activity of untested or even hypothetical compounds. In addition the insights generated by SAR studies can guide a researcher's attention to molecular features that are correlated with the biological activity of interest, thus confirming or contradicting mechanisms of action or suggesting further experiments. The techniques of SAR have been applied in the pharmaceutical field (drug design) and, to some extent, in agricultural chemicals. The methods are beginning to be applied to problems in the areas of chemical and genetic toxicity.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS (This Report)	21. NO. OF PAGES
RELEASE TO PUBLIC	UNCLASSIFIED	20
	20. SECURITY CLASS (This page) UNCLASSIFIED	22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

REPRODUCED BY
NATIONAL TECHNICAL INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

EPA/600/D-86/096

Chapter 7

Studies of Relationships between Molecular Structure and Biological Activity by Pattern Recognition Methods

Peter C. Jurs

INTRODUCTION

The attempt to rationalize the connections between the molecular structures of organic compounds and their biological activities comprises the field of structure-activity relations (SAR) studies. Correlations between structure and activity are important for the understanding and the development of pharmaceutical agents, agricultural agents, and chemical communicants (olfactory and gustatory stimulants) and for the investigation of chemical and genetic toxicity (mutagenic and carcinogenic potential). Practical importance attaches to this type of study because the potential exists for prediction of the biological activity of untested or even hypothetical compounds. In addition the insights generated by SAR studies can guide a researcher's attention to molecular features that are correlated with the biological activity of interest, thus confirming or contradicting mechanisms of action or suggesting further experiments. The techniques of SAR have been applied in the pharmaceutical field (drug design) and, to some extent, in agricultural chemicals. The methods are beginning to be applied to problems in the areas of chemical and genetic toxicity.

The most desirable way to develop the capability to predict the activity of untested compounds is to exploit a molecular level theory of action or mechanism. Unfortunately, this knowledge is not yet available for most classes of biologically active compounds and this approach is ordinarily blocked in problems of immediate interest. Thus one must use correlative methods to look for relationships between the molecular structures of tested compounds and their experimentally observed biological activity. Generally, one has available a set of compounds that have been tested in a standard bioassay and the results of these experiments. One can then use the methods of SAR to search for relationships between the structures and activities. This approach attempts to simultaneously take into account the entire process of uptake, transport, distribution, metabolism, cell penetration, binding, and excretion by correlating the structure of the administered compound with the final observed activity.

Several different approaches to the study of SAR have been reported. The most widely employed is the semi-empirical linear free energy (LFER) or extrathermodynamic model proposed by Hansch and co-workers (1-3), which is described elsewhere in this volume. This chapter will focus on an approach to SAR that brings together techniques of chemical structure information handling and pattern recognition in an attempt to find relationships between molecular structure and biological activity.

The fundamental premises involved in applying pattern recognition methods to SAR studies of chemicals with genetic toxicity are as follows:

Molecular structure and biological activity (genetic toxicity) are related.

The structures of compounds having genetic toxicity and compounds of similar structural classes that are nontoxic can be adequately represented by a set of molecular structure descriptors.

A relation can be discovered between the structure and activity by applying statistical and pattern recognition methods to a set of tested compounds.

The relation can be extrapolated to untested compounds.

The heart of the approach is finding a set of adequate descriptors for a particular data set consideration, that is, a set of descriptors for which a discriminating relation can be found.

The fundamental steps involved in performing an SAR study are shown in Fig. 1. The individual steps are as follows:

1 Identify, assemble, input, store, and describe a data set of structures for chemicals that have been tested for biological activity.

2 Develop computer-generated molecular descriptors for each of the members of the data set. The descriptors may be derived directly from the

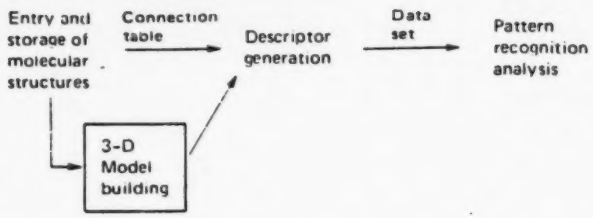

Figure 1 Flow chart of steps involved in structure-activity studies using chemical structure information handling and pattern recognition methods.

stored topological representations of the structures, or they may require the development of three-dimensional molecular models.

3 Using pattern recognition or statistical methods, develop classifiers to discriminate between active and inactive compounds based on the sets of molecular descriptors.

4 Test the predictive ability of these discriminants on compounds of unknown activity.

5 Systematically reduce the set of molecular structure descriptors employed to the minimum set sufficient to retain discrimination between the active and inactive compounds and to retain high predictive ability.

This approach to SAR studies has been implemented in a computer software system called ADAPT (*a*utomatic *d*ata *a*nalysis using *p*attern recognition *t*echniques) (4, 5). The ADAPT system currently consists of approximately 60 FORTRAN language routines that implement the steps necessary for an SAR study. It is designed to be run on a dedicated minicomputer or a time-sharing system, and it is highly interactive and must be guided by the user. The system is a set of tools that the user can apply to problems involving up to several hundred compounds at a time.

REPRESENTATION OF MOLECULAR STRUCTURE

The central problem of any correlative analysis is the choice of independent variables (here, molecular structure descriptors) that are used. The researcher's knowledge and intuition are brought to bear on the problem at hand through descriptor choice and refinement. Current hypotheses on mechanism of action, metabolism, transport, binding, etc., can be expressed through appropriately chosen descriptors. If a set of molecular structure descriptors that do contain the essential structural information relevant to the biological activity of interest is developed, then the analysis phase of the project will succeed. If an inadequate set of descriptors is analyzed, then no pattern recognition or

statistical procedure will unearth relationships. The search for adequate or good sets of descriptors forms one of the most challenging aspects of SAR studies.

A number of different approaches to descriptor development coexist in the SAR field. Historically, substituent constants—coding for structural differences among compounds in a homologous series—have enjoyed the most use (6). These descriptors include Hammett sigma constants for electronic effects, Taft steric parameters, π lipophilicity parameters, Verloop steric constants, and many others. This type of structural descriptor is fully described elsewhere in this volume and will not be stressed here. Electronic parameters resulting from quantum mechanical calculations have also been used in SAR work, either alone or in conjunction with other parameters (7). These too are described elsewhere in this volume.

The approach to structural representation to be described here involves the entry of storage of molecular topologies graphically, the modeling of the structures, and the calculation of the descriptors from these internal representations. This approach permits the study of large sets of compounds, heterogeneous sets of compounds, and the prediction of activity of unknown from only a sketch of the structure. Of course, if experimental data are found to be critical to the development of a structure-activity relationship, then such data can easily be included in the analysis.

The following paragraphs describe one approach to the development of structural descriptors for SAR studies.

Entry of Molecular Structures

The ADAPT computer system has as one of its components all the modules necessary to enter, modify, retrieve, and draw molecular structures of organic molecules. The routines allow the convenient, interactive entry of structures by sketching them on the screen of a graphics display terminal. This can be done in 30 s to several minutes per compound, depending on structural complexity. No special techniques beyond those used in sketching molecular structures on a blackboard are needed. Thus, structure files on the order of hundreds of compounds can be entered in reasonable amounts of time. The structure files are stored permanently on disc files for further processing. Information saved for each compound includes a compressed connection table, ring information, a list of associated numerical information, and identification number, the chemical name of the compound, and the two-dimensional coordinates of the atoms as entered (for possible redrawing later or for starting coordinates for modeling).

Molecular Mechanics Model Builder

The three-dimensional molecular model builder routine is used to derive information on the spacial conformation of molecules. A molecule can be

viewed as a collection of particles held together by simple harmonic or elastic forces. These forces can be defined by potential energy functions whose terms are the atom coordinates of the molecule. This function can then be minimized to obtain a strain-free three-dimensional model of the molecule. In the strain minimization section, the atom coordinates are systematically altered until a minimum is found in the strain or potential energy function. The strain function used is:

$$E_{strain} = E_{bond} + E_{angle} + E_{torsion} + E_{non\text{-}bond} + E_{stereo}$$

The first four terms of the function are commonly found in all molecular mechanics strain functions and are modified Hooke's law functions. The last term of the function has been added to assure the proper stereochemistry about an asymmetrical atom. An adaptive pattern search routine is used to minimize the strain energy because it does not require analytical derivatives. The amount of time necessary to obtain good molecular models depends on the number of atoms in the molecule, the initial strain of the molecule, and the degrees of freedom in the structure.

Descriptor Generation

The heart of SAR studies is the development of molecular structure descriptors. One of the major premises of the approach is that one can find an "adequate" set of descriptors to represent the compounds of interest. The existence of an adequate set of descriptors does not necessarily imply that they will be easily found. Thus descriptor development is the area in which chemists test their ingenuity most intensively, bringing to bear on the problem at hand all their insight and knowledge. It is in the area of descriptor development that the most difficult and most potentially rewarding parts of SAR research occur.

Fragment Descriptors These include counts of the number of atoms of each type, the number of bonds of each type, the number of basis rings, and the number of ring atoms.

Substructure Descriptors ADAPT has a substructure searching routine that can be used to develop descriptors. Each of the structures comprising a set of compounds under study is searched for the presence of the substructure of interest. If it is present, then the number of occurrences is computed. If not, then the descriptor is given the value of zero. The substructures to be used are problem dependent and must be found through the application of experience and outside knowledge by the researcher.

Environment Descriptors The information present in the fragment and substructure descriptors indicates the components of the molecular structure. However, the manner of interconnection is missing. Environment descriptors supply information about the connections by coding the immediate surroundings of substructures. To generate an environment descriptor, the molecule being coded is searched for the presence of the substructural fragment that forms the heart of the environment being sought. If no match is found, the descriptor is given the value of zero. If the substructure is found, then the descriptor is computed by performing a path one molecular connectivity calculation on the atoms comprising the substructure, as imbedded within the structure, and, in addition, the first nearest neighbor atoms. Thus, the value of the path one molecular connectivity represents the immediate surroundings of the substructure as imbedded within the molecule being coded.

Molecular Connectivity Descriptors The molecular connectivity (8) of a molecule is a measure of the branching of the structure. It is formed by summing contributions for each bond in the structure, where the contribution of each bond is determined by the connectivity of the atoms that are joined by that bond. This is the path one molecular connectivity. Higher order molecular connectivities can also be computed by considering all paths of length two, three, etc. These descriptors have been shown in several published reports to be correlated with a number of physicochemical parameters, such as partition coefficients and steric parameters.

Geometric Descriptors Once the molecules being studied have been modeled, then descriptors can be generated from these models. Geometrical descriptors that have been used in many SAR studies are the three principal moments of inertia of the molecules, the three ratios of these three principal moments, and the molecular volume. Many other geometrical descriptors can be envisioned, and the development of new descriptors is an active area of research.

Path and Path Environment Descriptors Molecular structures can be viewed in the abstract as graphs, in the mathematical sense. Graphs have a number of properties that can be readily computed, among them paths. A *path* is a sequence of connected atoms in which no atom appears more than once. The tracing of paths in highly connected graphs, e.g., polycyclic aromatic hydrocarbons, is tedious but can be done. A method for finding all the paths contained in a graph has been published, with a discussion of the chemical utility of this measure of bonding complexity (9). A descriptor generation routine has been implemented and interfaced to ADAPT using this concept. The routine calculates the total number of paths of all lengths contained within the structure being coded.

The path environment descriptor is similar to the molecular connectivity environment descriptor. A substructure is matched against the structure being coded. If it is present, then a pseudomolecule is formed from the substructure atoms and their first nearest neighbor atoms. Then the number of paths originating from the set of atoms forming the pseudomolecule is found. The maximum path length to be considered is input by the user. The descriptor for a particular structure consists of the number of paths originating from the substructure atoms and their first nearest neighbors. This environment can be made sensitive to the immediate surroundings of the substructure by making the maximum path length to be considered very small, e.g., two or three. Alternatively, it can be made sensitive to the overall structure of the molecule by making the maximum path length larger.

Partition Coefficient Estimation The log P, the logarithm of the partition coefficient of a compound between an aqueous phase and a model lipid phase, is highly correlated with various types of biological activities. Its inclusion as a descriptor allows the system to have access to information regarding transport and partioning that may not be as clearly represented in the other descriptors. The calculation of log P can be done using a "constructionist approach" (6). This formulation starts with a set of fundamental fragments the values of which are summed with appropriate weightings. Corrections are added to refine the calculation if necessary. The equation used is as follows:

$$\log P = \sum_{i=1}^{n} a_i f_i + \sum_{j=1}^{n} c_j$$

where f_i is the fragment constant for the ith fragment, a_i is the number of occurrences of the ith fragment, and c_j is the jth correction factor. The fundamental assumption of this approach is that hydrophobicity is an additive-constitutive property of molecules.

Recently, this fragment constant method has been implemented in a routine and interfaced to ADAPT (10). Thus, the estimated log P value for a molecule can be used as a descriptor along with all the other descriptor types.

Descriptor Manipulation Routines In addition to the individual descriptor generation routine, the ADAPT system contains several support routines. There is a general maintenance and interrogation routine for the descriptor files that allows the user to review descriptor values by listing them and to delete unwanted descriptors. Another routine allows the user to input from cards descriptors obtained from any source so that they can be studied along with the computer-generated descriptors. A manipulation routine allows the user to alter previously generated or stored descriptors to form new

descriptors. Addition, subtraction, multiplication, division, exponentiation, logarithmic transformation, autoscaling, and inversion are allowed.

The development of adequate sets of descriptors for the compounds forming a data set comprises the most difficult part of SAR research. With an adequate set of descriptors, the analysis portion of the study is relatively straightforward. With a set of descriptors that is inadequate, one has no choice but to keep searching for better descriptors. Thus, descriptor development for a particular data set can consume quite a lot of time and can be a trial-and-error operation.

A key goal in descriptor development is to find descriptors that are related to the mode of action and the mechanism or metabolism responsible for the biological activity of interest. This will insure the most meaningful representation of the structures under investigation.

PATTERN RECOGNITION ANALYSIS

Once each compound in a data set under investigation has been represented by a set of descriptors, then the analysis phase of the structure-activity study begins. In this chapter, pattern recognition analysis methods will be discussed. Other alternative methods of data analysis are described elsewhere in this volume.

Pattern recognition is a subfield of artificial intelligence developed largely by electrical engineers and computer scientists. It comprises a set of parametric and nonparametric techniques used to study data sets that do not conform to well-characterized probability density functions. A voluminous literature describes the field (11–13).

Most of the pattern recognition methods share a set of common properties, including the method of representation of the data. Each compound is represented by a vector

$$\mathbf{X} = (x_1, x_2, \ldots, x_n)$$

wherein each individual component of the pattern vector, $\mathbf{x}_j$, is the value for the jth descriptor. To reference the ith compound in a data set, the notation $\mathbf{X}_i$ is used. An equivalent view of this representation is to consider each component to be represented by a point in n-dimensional space.

For a given compound, which is represented by a given point, the value of each coordinate is just the numerical value for one of the molecular structure descriptors comprising the representation. The expectation is that the points representing compounds of common biological activity (e.g., toxic compounds) will cluster in one limited region of the space separate from those compounds of another biological activity (e.g., nontoxic compounds), which will cluster elsewhere. The clusters are regions of high local density relatively

far apart from each other. Pattern recognition consists of a set of methods for investigating data represented in this manner to assess the degree of clustering and general structure of the data space.

Preprocessing

Preprocessing operations alter the form of the representations of the compounds and include operations such as scaling, normalization, and other mathematical transformations designed to make the analysis easier at a later stage.

Scaling and normalization are required to convert the given units from different descriptors to a compatible form. For example, molecular weights ranging from 300–500 should not be studied simultaneously with log P values ranging from 3–5. In such cases the larger valued descriptors could dominate those of smaller value and lead to numerical instabilities in the pattern recognition programs. One preprocessing method that is quite useful in correcting this type of imbalance is autoscaling.

This method simultaneously scales and normalizes the data by translating it so that the mean is zero and normalizing it so that each descriptor has a standard deviation of S. The equations describing this process are given below.

$$x'_{ij} = \frac{S(x_{ij} - x_j)}{\sigma_j}$$

$$x_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

$$\sigma_j^2 = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - x_j)^2$$

The x_{ij} are the elements of the matrix representation of the data formed from the column vectors representing each data point, and the x'_{ij} are the elements of the data matrix resulting from the scaling operation. Autoscaling can be looked upon as "squaring up the space," since it places the data inside a hypercube. Although autoscaling alters the spread of the data, it does not alter the number of features or the basic geometry of the clustering.

Mapping and Display

One powerful method for analyzing the structure of a data set (here, a set of compounds) is to map the points from the high-dimensional space to a plane for direct viewing. It has long been recognized that people are adept at

recognizing patterns in two-dimensional plots, and this has been exploited in many pattern recognition display methodologies. A number of useful and imaginative graphic techniques exist for directly displaying multivariate data in just one or two dimensions (14). Their use has been largely neglected in the physical sciences in favor of more common rectangular plots and histograms. However, techniques such as metroglyphs, linear and circular data profiles, Andrews plots, and Chernoff faces can be valuable tools for visualizing relationships between observations, for identifying outliers, and for classification purposes (15, 16).

A method that combines both mapping and display operations along with feature selection operations is principal components analysis, also known as the Karhunen-Loeve transformation (11). This involves taking the eigenvalues and eigenvectors of the variance-covariance matrix of a data set and then using the eigenvectors to perform a linear rotation. The multidimensional data set can be plotted in the two dimensions determined by the two principal eigenvectors and displayed for visual analysis. Alternatively, the data set can be rotated by multiplication with any desired number of eigenvectors to reduce the dimensionality of the data set. This operation commonly involves using just enough of the available eigenvectors to retain 95 to 99 percent of the total sum of the values. An advantage of this feature selection procedure is that the number of descriptors per compound is reduced, thus making further analysis more convenient in some ways. A disadvantage is that the individual descriptors resulting from this transformation are mixtures of the original descriptors, thus complicating analysis of later results.

Another mapping strategy is to attempt to find a two-dimensional display in which each point in the original data set is represented by a point in the two-dimensional array of points. Of course, a certain amount of error in the mapping is to be expected. The literature contains reports of many approaches to perform these nonlinear mapping operations (17-19). Iterative nonlinear mapping routines have been implemented using the following error function:

$$E = \frac{1}{\sum_{i<j} d_{ij}^*} \sum_{i<j} \frac{(d_{ij}^* - d_{ij})^2}{d_{ij}^*}$$

where d_{ij}^* is the euclidean distance between points i and j in the original n-dimensional space and d_{ij} is the two-dimensional interpoint distances. The method of steepest descent can be used to minimize E. The starting estimates may be chosen by the user: they could be the results of a previous principal components analysis, for example. Many other starting configuration options are also available. The results of the mapping routine could be presented graphically for visual analysis.

Classification by Discriminants

Commonly in SAR studies, the compounds available for study have been placed in a few discrete categories such as "active" versus "inactive." After the compounds have been represented by molecular structure descriptors and expressed as points in a high-dimensional space, then the question becomes, "Are the two sets of points representing compounds separable from one another?" That is, can a discriminant surface be placed between the point sets so that all points representing the active compounds fall on one side of the surface and all of the points representing inactive compounds fall on the other side of the surface? The search for such discriminants comprises the operation of classification and discriminant development.

Two different approaches to the development of discriminants are feasible: parametric and nonparametric methods. Parametric methods of pattern recognition attempt to find classification surfaces or clustering definitions based on statistical properties of the members of one or both classes of points. For example, bayesian classification surfaces are developed using the mean vectors for the members of the classes and the covariance matrices for the classes. If the statistical properties cannot be calculated or estimated, then nonparametric methods are used. Nonparametric methods attempt to find clustering definitions or classification surfaces by using the data themselves directly, without computing mean vectors, covariance matrices, etc. Examples of nonparametric methods would include error-correction feedback linear learning machines (threshold logic units or perceptrons) and simplex optimization methods of searching for separating classification surfaces.

Parametric Pattern Recognition Programs The parametric pattern recognition programs use the mean vectors and covariance matrices (or other statistical measures) of the two classes into which the patterns fall as their basis for development of discriminants (11).

One parametric approach implements a quadratic discriminant function using the Bayesian theorem. The equation for the discriminants is as follows.

$$d_k(X) = \ln p_k - \tfrac{1}{2} \ln |C_k| - \tfrac{1}{2} [(X - \mathbf{m}_k)^T C_k^{-1} (X - \mathbf{m}_k)] \qquad k = 1,2$$

where p_k is the a priori probability for class k, C_k is the covariance matrix for class k, and $\mathbf{m}_k$ is the mean vector for class k. This discriminant assumes a multivariate normal distribution of the data. A pattern, $\mathbf{X}_i$, is put in the class for which $d_k(\mathbf{X}_i)$ is greatest.

If the assumption is made that the covariance matrices for the two classes are the same, then the discriminant function developed using the Bayes theorem and the multivariate normal assumption simplifies from that above to the following:

$$s = d_1(X) - d_2(X) = \ln p_1 - \ln p_2 + X'X^{-1}(m_1 - m_2) - \tfrac{1}{2} m_1' C^{-1} m_1 + \tfrac{1}{2} m_2' C^{-1} m_2$$

where s is greater than zero for one class and less than zero for the other class.

Another parametric approach implements a discriminant function by the method commonly called linear discriminant function analysis (LDFA). It is nearly identical to the linear bayesian discriminant, except that instead of using the covariance matrix, the sum of cross-products matrix is used. Results obtained with LDFA are ordinarily very similar to those obtained using a Bayes routine.

Nonparametric Pattern Recognition Programs The nonparametric pattern recognition programs develop their discriminants using the training set of patterns to be classified rather than statistical measures of their distributions.

The k nearest neighbor method is conceptually the simplest. An unknown pattern is assigned to the class to which the majority of its nearest neighbors belong. The metric that is used to determine proximity is ordinarily the euclidean metric, but any metric can be used. Another widely used nonparametric method is the linear learning machine or perceptron (11). The algorithm is heuristic. A decision surface is initialized either arbitrarily or using the result from another linear discriminant development routine. Then the learning machine classifies one member of the training set at a time. When the current discriminant function correctly classifies a pattern, the discriminant is left unchanged. However, when an incorrect classification is made, then the discriminant is altered in such a way that the error just committed is eliminated. The learning machine continues to classify the members of the training set repeatedly until no errors are committed or until the user terminates the routine. This algorithm for development of linear discriminant functions has the desirable property that if a training set of patterns is separable into the two classes, i.e., if a solution exists, then this method will find a solution. This routine is the one ordinarily used to train a set of slightly different weight vectors to be used for the variance method of nonparametric feature selection method.

Another nonparametric method develops a linear discriminant function through an iterative least squares approach (12). The following function is minimized:

$$Q = \sum_{i=1}^{m} [Y_i - F(s_i)]^2$$

where m is the total number of patterns in the training set, Y_i equals +1 for

one pattern class and −1 for the other class, s_i is the dot product of the weight vector with the ith pattern vector, and $F(s_i)$ is the hyperbolic tangent function of s_i. The hyperbolic tangent function has the values of positive one for all positive values of s_i and negative one for all negative values of s_i. Therefore, minimizing Q is equivalent to minimizing the number of incorrect classifications. The function is nonlinear in the independent variables and thus cannot be solved directly, so an iterative algorithm is employed. This routine has been found to be particularly useful for dealing with data that are not separable into the two classes.

Recently, Moriguchi and co-workers (20) have reported an adaptive least squares method for the development of a multiclass discriminant function, and they have applied the method to a structure-activity problem.

The search for a good discriminant that separates two clusters of points in an n-dimensional space from one another can be formulated as a linear programming problem (21). The objective function whose value is to be minimized can be defined as the fraction of the training set of patterns that are incorrectly classified. Because two different discriminants can easily have the same classification power, it has been suggested that a secondary objective function be defined as the sum of the distances from the discriminant plane to the misclassified points. The secondary objective function is invoked only if the primary function is equal for two different discriminants being compared. The simplex optimization procedure described above has been applied to a number of chemical problems.

Once discriminants have been found that do separate the data set into the appropriate subsets, then these discriminants can be used to assess validity. This is usually done by a round-robin procedure involving leaving out a small number of data set members to act as "unknown" compounds. When available, true unknowns can also be input into the system for prediction of activity.

Clustering Methods

A subset of pattern recognition methods consists of clustering methods that attempt to determine structural characteristics of a set of data by organizing the data into subgroups or clusters. These methods are known as "unsupervised" because they attempt to define clusters solely on the basis of criteria derived from the data themselves (as opposed to finding clusters after knowing which data points belong to which categories). The most widely used clustering method is hierarchical clustering. This method works by measuring the distances between all pairs of points, identifying the closest pair, combining them into a single point at the midpoint, recalculating the distances from this new point to every other point in the data set, finding the nearest pair, combining them, etc. The process is repeated until all the points have

been combined. The resulting structure can be displayed as a dendogram, which shows at what degree of similarity each pair of points was combined. The overall purpose is usually to simplify a data matrix that is too extensive for direct analysis. The product of hierarchical clustering analysis is a dendogram or treelike structure that shows visually the hierarchy of similarity relationships inherent in the original data. The techniques are independent of the nature of the material to be classified, and they have been applied to a wide variety of problems drawn from engineering, the social sciences, the physical sciences, medicine, and biology (19, 22–24). Clustering routines must be considered to be an exploratory tool (25), and the absolute validity of dendograms generated may be less important than the insights and suggestions generated regarding the data structure.

Simca Pattern Recognition

A pattern recognition method has been developed by Wold and co-workers called SIMCA (26–28). SIMCA is a set of procedures with which each class of pattern recognition problems is separately modeled with a principal components model. Unknowns to be classified are fitted to the class models, and the classifications are made according to the statistics of the fits. An unknown pattern is assigned a probability for each class: if all probabilities are low, then the pattern may be an outlier belonging to none of the original classes.

PATTERN RECOGNITION IN STRUCTURE-ACTIVITY RELATIONSHIP STUDIES

The methods of pattern recognition have been applied to the problem of searching for correlations between molecular structure and biological activity. The types of biological activity studied include pharmaceutical activity (drug design), agricultural chemicals, chemical communicants (olfactory stimulants), and toxicity (chemical, mutagenic, and carcinogenic activity). Applications of pattern recognition to drug design have been reviewed by Kirschner and Kowalski (29) and briefly discussed by Varmuza (13). A book describing research on pattern recognition in SAR has appeared (5).

A few papers have appeared in the past several years dealing with specific structure-toxicity studies. Each of these will be described briefly in the following paragraphs.

A paper by Norden et al. (30) reported a study of 32 PAH using the SIMCA pattern recognition methodology. Each PAH was characterized by 23 variables. These included 15 theoretical variables (electronic descriptors, molecular orbital energies, charges of radical cations, symmetry descriptors, etc.) and eight measured variables (six spectroscopic values extracted from absorption spectra and two other spectroscopic variables). A procedure was

developed to classify the 32 PAH as carcinogenic or noncarcinogenic and to provide a quantitative prediction about the level of activity of the active compounds. Validation studies showed the classification rate to be significantly better than chance. The correlation coefficient between predicted and observed activity during the validation test was 0.793.

A paper by Dunn and Wold (31) reported a study of 33 4-nitroquinoline-1-oxides using the SIMCA pattern recognition methodology. Each compound was characterized by 43 variables: six positions of possible substitution with six physicochemical substituent constants each (π, MR, σ_m, σ_p, L, and B_4), two positions with three additional substituent constants used (B_1, B_2, and B_3), and log P. A four-component principal components model was derived for the carcinogens, which was 82 percent successful in predicting the carcinogenic potential of the compounds. Validation of the classification procedure developed showed a highly statistically significant relationship to exist.

A paper by Jurs and co-workers (32) reported a study of 209 heterogeneous compounds drawn from more than 12 structural classes. Twenty-six calculated molecular structure descriptors were used; these supported a linear discriminant function capable of separating 192 compounds into the carcinogen and noncarcinogen classes. Validation of the linear discriminant achieved a classification success rate of 90 percent for carcinogens and 78 percent for noncarcinogens.

A paper by Chou and Jurs (33) described studies of 153 *N*-nitrosamines and *N*-nitrosamides. A set of 15 calculated molecular structure descriptors was found that supported a linear discriminant function capable of separating 116 carcinogens from 28 noncarcinogens. Randomized validation testing produced a classification ability of 93 percent for carcinogens and 85 percent for noncarcinogens. The inclusion of two electronic descriptors representing the electronic properties of the N-NO group and the α-carbon atom implicitly supports the α-hydroxylation hypothesis of action of *N*-nitroso compounds.

A paper by Yuan and Jurs (34) reported the application of pattern recognition to the study of carcinogenic potential of PAH compounds drawn from 16 structural subclasses. Each PAH was represented by 28 calculated molecular structure descriptors. Linear discriminants were found that could classify 191 of the 200 PAH as carcinogens versus noncarcinogens. Validation testing produced 91 percent correct classifications.

Dunn and Wold (35) have reported a study of *N*-nitroso compounds, including nitrosamines, *N*-nitrosoureas, and *N*-nitrosourethans. Each compound was represented by six substituent constants for two sites (Rekker lipophilicity constant, Taft's σ^* and E_s, *MR*, Verloop's steric constants L and B_4). The active compounds were subdivided into three classes and models were developed for each of the three: a two-component model correctly classified 23 of 27 in class 1; a one-component model correctly classified 7 of 9 in class

2: a three-component model correctly classified all 14 compounds in class 3. Thus 44 of 50 or 88 percent of the active compounds were classified correctly overall. All of the inactive compounds, except one, were classified as members of none of the three classes.

A paper by Yuta and Jurs (36) describes studies of aromatic amines using pattern recognition methods. The 157 aromatic amine compounds were divided into subsets according to active site (breast, ear duct, liver, other, all sites), route of administration (oral or other), and activity (carcinogen or noncarcinogen). Sets of calculated molecular structure descriptors were generated that could support linear discriminant functions able to separate sets of active carcinogens from inactive compounds. Prominent among the important structural descriptors were those coding sizes and shapes of the amines. The pattern recognition results were not strongly affected by differences in active site, and the study showed that mixed data could be used in computer-assisted structure-carcinogenicity studies.

A study by Whalen-Pedersen and Jurs (37) describes the results of a study of relationships between molecular structures of polycyclic aromatic compounds (PAC) and capillary column gas chromatographic retention indexes. Experimentally measured retention indexes, on a linear scale that involves interpolation between PAH standards, for 207 PAC were found to be well explained by a four-variable linear regression model. The four calculated structural descriptors were molecular volume, number of basis rings, path one molecular connectivity, and the number of nitrogen plus oxygen atoms. For 207 compounds the correlation coefficient obtained was 0.988 and standard error was 12.4 compared to the range of the dependent variable, which was 197 to 504.

REFERENCES

1 Hansch, C. Recent advances in biochemical QSAR. In N. R. Chapman and J. Shorter (eds.), *Advances in Linear Free Energy Relationships*. New York: Plenum Press (1979).

2 Martin, Y. C. *Quantitative Drug Design*. New York: Dekker (1978).

3 Unger, S. H. Consequences of the Hansch paradigm for the pharmaceutical industry. In E. J. Ariëns (ed.), *Drug Design*, vol. IX, pp. 48–119. New York: Academic Press (1980).

4 Stuper, A. J., W. E. Brugger, and P. C. Jurs. A computer system for structure-activity studies using chemical structure information handling and pattern recognition techniques. *A.C.S. Symp. Ser.* 52:165–191 (1977).

5 Stuper, A. J., W. E. Brugger, and P. C. Jurs. *Computer Assisted Studies of Chemical Structure and Biological Function*. New York: Wiley-Interscience (1979).

6 Hansch, C., and A. Leo. *Substituent Constants for Correlation Analysis in Chemistry and Biology*. New York: Wiley-Interscience (1979).

7 Christoffersen, R. E. Quantum pharmacology: Recent progress and current status. In E. C. Olsen and R. E. Christoffersen (eds.), *Computer Assisted Drug Design*. Washington, D.C.: American Chemical Society (1979).

8 Kier, L. B., and L. H. Hall. *Molecular Connectivity in Chemistry and Drug Research*. New York: Academic Press (1976).

9 Randic, M., G. M. Brissey, R. B. Spencer, and C. L. Wilkins. Search for all self-avoiding paths for molecular graphs. *Comput. Chem.* 3:5–13 (1979).

10 Chou, J. T., and P. C. Jurs. Computation of partition coefficients from molecular structures by a fragment addition method. In S. H. Yalkowsky, A. A. Sinkula, and S. C. Valvani (eds.), *Physical Chemical Properties of Drugs*, pp. 163–199. New York: Dekker (1980).

11 Tou, J. T., and R. C. Gonzalez. *Pattern Recognition Principles*. Reading, Mass.: Addison-Wesley (1974).

12 Jurs, P. C., and T. L. Isenhour. *Chemical Applications of Pattern Recognition*. New York: Wiley-Interscience (1975).

13 Varmuza, K. *Pattern Recognition in Chemistry*. Berlin: Springer-Verlag (1980).

14 Gnanadesikan, R. *Methods for Statistical Data Analysis of Multivariate Observations*. New York: Wiley, (1977).

15 Mezzich, J. E. *Graphical Representation of Multivariate Data*. New York: Academic Press (1978).

16 Chernoff, H. *The Use of Faces to Represent Points in n-Dimensional Space Graphically*. Technical Rep. No. 71, Department of Statistics, Stanford University, Stanford, CA (1971).

17 Duda, R. O., and P. E. Hart. *Pattern Classification and Scene Analysis*. New York: Wiley (1973).

18 Sammon, J. W. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* C-18:401 (1979).

19 Fukunaga, K. *Introduction to Statistical Pattern Recognition*. New York: Academic Press (1972).

20 Moriguchi, I., K. Komatsu, and Y. Matsushita. Adaptive least-squares method applied to structure-activity correlation of hypotensive N-alkyl-N″-cyano-N″-pyridylguanidines. *J. Med. Chem.* 23:20–26 (1980).

21 Ritter, G. L., S. R. Lowry, C. L. Wilkins, and T. L. Isenhour. Simplex pattern recognition. *Anal. Chem.* 47:1951 (1975).

22 Everitt, B. *Cluster Analysis*. New York: Wiley (1974).

23 Sokal, R. R., and P. H. A. Sneath. *Principles of Numerical Taxonomy*. San Francisco: Freeman (1963).

24 Hartigan, J. A. *Clustering Algorithms*. New York: Wiley (1975).

25 Dubes, R., and A. K. Jain. Validity studies in clustering methodologies. *Pattern Recog.* 11:235 (1979).

26 Wold, S. Pattern recognition by means of disjoint principal components models. *Pattern Recog.* 8:127–139 (1976).

27 Wold, S., and M. Sjostrom. SIMCA: A method for analyzing chemical data in terms of similarity and analogy. *A.C.S. Symp. Ser.* 52:243–282 (1977).

28 Dunn, W. J., III, and S. Wold. Relationships between chemical structure and biological activity modeled by SIMCA pattern recognition. *Bioorganic Chem.* 9:505–523 (1980).

29 Kirschner, G. L., and B. R. Kowalski. The application of pattern recognition to drug design. In E. J. Ariëns (ed.), *Drug Design*, vol. 8, pp. 73–131. New York: Academic Press (1979).

30 Norden, B., U. Edlund, and S. Wold. Carcinogenicity of polycyclic aromatic hydrocarbons studied by SIMCA pattern recognition. *Acta Chem. Scand.* [B]32:602–608 (1978).

31 Dunn, W. J., and S. Wold. A structure-carcinogenicity study of 4-nitroquinoline 1-oxides using the SIMCA method of pattern recognition. *J. Med. Chem.* 21:1001–1007 (1978).

32 Jurs, P. C., J. T. Chou, and M. Yuan. Computer assisted structure-activity studies of chemical carcinogens. A heterogeneous data set. *J. Med. Chem.* 22:476–483 (1979).

33 Chou, J. T., and P. C. Jurs. Computer assisted structure-activity studies of chemical carcinogens. N-nitroso compounds. *J. Med. Chem.* 22:792–797 (1979).

34 Yuan, M., and P. C. Jurs. Computer assisted structure-activity studies of chemical carcinogens. Polycyclic aromatic hydrocarbons. *Toxicol. Appl. Pharmacol.* 52:294–312 (1980).

35 Dunn, W. J., III, and S. Wold. The carcinogenicity of N-nitroso compounds: A SIMCA pattern recognition study. *Bioorganic Chem.* (in press).

36 Yuta, K., and P. C. Jurs. Computer assisted structure-activity studies of chemical carcinogens. Aromatic amines. *J. Med. Chem.* (in press).

37 Whalen-Pedersen, E. K., and P. C. Jurs. Computer assisted structure-activity studies of polycyclic aromatic hydrocarbons. In *Fifth International Symposium on Polynuclear Aromatic Hydrocarbons.* Columbus, Ohio: Battelle Press (in press).